# Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia

Zhou, Zhemin; Lundstrøm, Inge Kristine Conrad; Tran-Dien, Alicia; Duchêne, Sebastián; Alikhan, Nabil-Fareed; Sergeant, Martin J.; Langridge, Gemma; Fotakis, Anna Katerina; Nair, Satheesh; Stenøien, Hans K.; Hamre, Stian S.; Casjens, Sherwood; Christophersen, Axel; Quince, Christopher; Thomson, Nicholas R.; Weill, Francois-Xavier; Ho, Simon Y. W.; Gilbert, Tom; Achtman, Mark

*Published in:*
Current Biology

*DOI:*
10.1016/j.cub.2018.05.058

*Publication date:*
2018

*Document version*
Publisher's PDF, also known as Version of record

Report

Current Biology

# Pan-genome Analysis of Ancient and Modern *Salmonella enterica* Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia

## Highlights

- *Salmonella enterica* aDNA sequences were found within 800-year-old teeth and bone
- The invasive Para C lineage was defined from 50,000 modern *S. enterica* genomes
- The Para C lineage includes Ragna, the aDNA genome, and human and swine pathogens
- Only few genomic changes occurred in the Para C lineage over its 3,000-year history

## Authors

Zhemin Zhou, Inge Lundstrøm, Alicia Tran-Dien, ..., Simon Y.W. Ho, M. Thomas P. Gilbert, Mark Achtman

## Correspondence

zhemin.zhou@warwick.ac.uk (Z.Z.), tgilbert@snm.ku.dk (M.T.P.G.), m.achtman@warwick.ac.uk (M.A.)

## In Brief

Zhou et al. reshape our understandings of the origins of an invasive bacterial pathogen, *Salmonella enterica* Paratyphi C, by combining a reconstructed pan-genome from an 800-year-old skeleton in Norway with 221 genomes from modern bacteria.

Zhou et al., 2018, Current Biology *28*, 2420–2428
August 6, 2018 
https://doi.org/10.1016/j.cub.2018.05.058

CellPress

CellPress

Current Biology
Report

# Pan-genome Analysis of Ancient and Modern *Salmonella enterica* Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia

Zhemin Zhou,[1,*] Inge Lundstrøm,[2] Alicia Tran-Dien,[3] Sebastián Duchêne,[4] Nabil-Fareed Alikhan,[1] Martin J. Sergeant,[1] Gemma Langridge,[5,10] Anna K. Fotakis,[2] Satheesh Nair,[5,11] Hans K. Stenøien,[6] Stian S. Hamre,[7] Sherwood Casjens,[8] Axel Christophersen,[6] Christopher Quince,[1] Nicholas R. Thomson,[5] François-Xavier Weill,[3] Simon Y.W. Ho,[9] M. Thomas P. Gilbert,[2,6,*] and Mark Achtman[1,12,*]

[1]Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK
[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark
[3]Unité des Bactéries Pathogènes Entériques, Institut Pasteur, Paris, France
[4]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3010, Australia
[5]Wellcome Trust Sanger Institute, Cambridge, UK
[6]NTNU University Museum, N-7491 Trondheim, Norway
[7]Department of Archaeology, History, Cultural Studies and Religion, University of Bergen, Post Box 7805, 5020 Bergen, Norway
[8]Pathology Department, University of Utah School of Medicine, Salt Lake City, UT 84112, USA
[9]School of Life and Environmental Sciences; University of Sydney, Sydney NSW 2006, Australia
[10]Present address: Molecular Microbiology, Norwich Medical School, University of East Anglia, Norwich, UK
[11]Present address: Gastrointestinal Bacteria Reference Unit, Public Health England, London, UK
[12]Lead Contact
*Correspondence: zhemin.zhou@warwick.ac.uk (Z.Z.), tgilbert@snm.ku.dk (M.T.P.G.), m.achtman@warwick.ac.uk (M.A.)
https://doi.org/10.1016/j.cub.2018.05.058

## SUMMARY

*Salmonella enterica* serovar Paratyphi C causes enteric (paratyphoid) fever in humans. Its presentation can range from asymptomatic infections of the blood stream to gastrointestinal or urinary tract infection or even a fatal septicemia [1]. Paratyphi C is very rare in Europe and North America except for occasional travelers from South and East Asia or Africa, where the disease is more common [2, 3]. However, early 20$^{th}$-century observations in Eastern Europe [3, 4] suggest that Paratyphi C enteric fever may once have had a wide-ranging impact on human societies. Here, we describe a draft Paratyphi C genome (Ragna) recovered from the 800-year-old skeleton (SK152) of a young woman in Trondheim, Norway. Paratyphi C sequences were recovered from her teeth and bones, suggesting that she died of enteric fever and demonstrating that these bacteria have long caused invasive salmonellosis in Europeans. Comparative analyses against modern *Salmonella* genome sequences revealed that Paratyphi C is a clade within the Para C lineage, which also includes serovars Choleraesuis, Typhisuis, and Lomita. Although Paratyphi C only infects humans, Choleraesuis causes septicemia in pigs and boar [5] (and occasionally humans), and Typhisuis causes epidemic swine salmonellosis (chronic paratyphoid) in domestic pigs [2, 3]. These different host specificities likely evolved in Europe over the last ~4,000 years since the time of their most recent common ancestor (tMRCA) and are possibly associated with the differential acquisitions of two genomic islands, SPI-6 and SPI-7. The tMRCAs of these bacterial clades coincide with the timing of pig domestication in Europe [6].

## RESULTS AND DISCUSSION

According to historical records [7], humans have long been afflicted by bacterial infections, yet genomic analyses of extant bacterial pathogens routinely estimate a tMRCA of no more than a few centuries [8]. In general, evolutionary trees contain a stem group, which may include lineages that are now rare or extinct, as well as the crown group of extant organisms. Historical reconstructions based only on the crown group ignore the older sub-lineages in the stem group and thereby provide an incomplete picture of the older evolutionary history of the pathogen. In contrast, analyses of ancient DNA (aDNA) can shed light on additional millennia of bacterial pathogen evolution that occurred prior to the origin of the crown group [9, 10]. We therefore searched for ancient bacterial lineages by scanning metagenomic sequences from teeth and long bones of 33 skeletons who were buried between 1100 and 1670 CE in Trondheim, Norway [11] (Figures 1A and 1C).

SK152 is the skeleton of a 19- to 24-year-old woman of 154 ± 3 cm height who was buried in 1200 ± 50 CE, according to archaeological investigations [11]. Calibrated radiocarbon ($^{14}$C) dating of two teeth estimated her burial as 100–200 years earlier (Figure S1B); this minor discrepancy may reflect the reservoir effect on radiocarbon dating from a predominant diet of fish products [13]. Based on $\delta^{18}O_{carbon}$ isotopic measurements from her first and third molars, this woman likely migrated from the northernmost inland areas of Scandinavia or Northwest

 

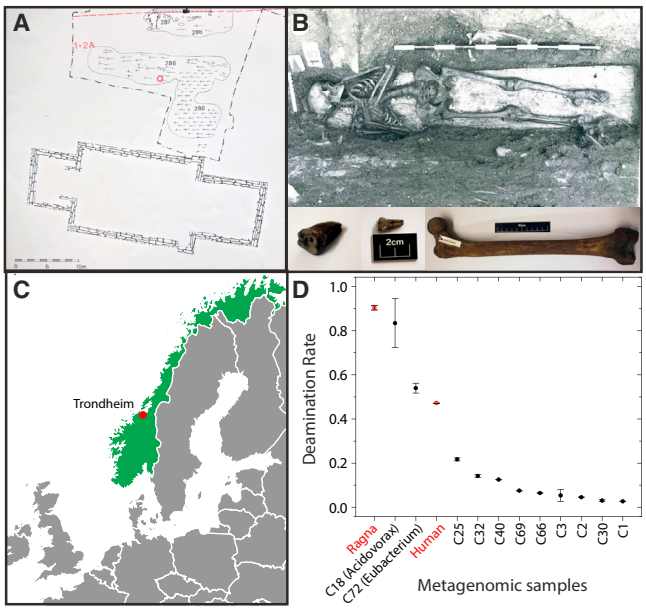


**Figure 1. Geographic, Archaeological, and Metagenomic Features of Skeleton SK152**

(A) Excavation site (Folkebibilotekstomten, 1973–1985) of the church cemetery of St. Olav in Trondheim, Norway. The burial location of SK152 (red circle) belongs to a building phase that has been dated archaeologically [11] to 1200 CE (range 1175–1225).

(B) Entire skeleton (top) and femoral long bone plus two teeth from which *Salmonella* DNA was extracted (bottom).

(C) Map of Europe surrounding Norway (green) and location of Trondheim (red).

(D) Deamination rate for metagenomic reads in the *Salmonella* Paratyphi C Ragna genome, human DNA and 11 single genome assemblies (Cxx) identified by Concoct [12]. C18 (*Acidovorax*) and C72 (*Eubacterium*) show high levels of deamination rates, as do reads from humans or Ragna, while the other assemblies have low levels and likely represent modern environmental bacteria. Data: Mean plus error bars showing standard deviations. See also Figure S2 for properties of aDNA reads.

Russia during her childhood and arrived in Trondheim by her early teens [14] (Figure S1A). The skeleton rested on a wooden plank (a symbolic coffin) in a grave filled and covered with anoxic, acidic, and waterlogged wood chips and soil (Figure 1B).

## Metagenomic Reads Associated with SK152

We identified 266 *Salmonella enterica* sequence reads in one tooth from skeleton SK152 and many more *Salmonella* sequences from additional libraries from SK152 teeth and bone, but not from dental calculus (Table 1). We attempted to reconstruct this *Salmonella* genome (henceforth designated "Ragna") from the SK152 sequence libraries. Concoct [12] was able to reconstitute 11 near-complete microbial genomes from those data (Figure 1D). Nine of these genomes likely reflect recent soil contamination because their 5′-single-stranded deamination rates were low [15], with DNA damage < 0.22, versus values of 0.47 for human DNA and 0.9 for Ragna.

The two other assembled genomes exhibited high levels of DNA damage and are thus likely to have been endogenous to this corpse since burial. Genome C72, from a novel species of *Eubacterium*, was found almost exclusively in dental calculus, and these bacteria may have been a component of a biofilm associated with periodontal disease, as are other *Eubacterium* species [16]. Genome C18 belongs to *Acidovorax*, which is associated with plant pathogens [17] and may have been introduced with the wood chips that covered the skeleton. We therefore decided to reconstruct the Ragna genome by read mapping against its close relatives within *S. enterica*.

## The Para C Lineage

Identifying close relatives of the Ragna genome required an overview of the genetic diversity of *S. enterica* subspecies *enterica.* To this end, we inferred phylogenetic trees of 2,964 genomes that represented the diversity of 50,000 strains of *S. enterica* in EnteroBase [18]. These contained 711,009 single-nucleotide polymorphisms (SNPs) in 3,002 core genes (2.8 Mb). A maximum-likelihood tree of the concatenated core genes revealed the existence of multiple, discrete lineages (Figure 2A). The initial sequence reads from Ragna were most closely related to one of these lineages, the "Para C lineage." The Para C lineage is comprised of monophyletic clades of serovars Paratyphi C, Choleraesuis, and Typhisuis, which were already known to be related by lower resolution analyses [3], plus one genome of the extremely rare serovar Lomita (Figure 2). Paratyphi C only infects humans, but serovar Choleraesuis is associated with septicemia in swine (and occasionally humans) and Typhisuis is associated with epidemic swine salmonellosis (chronic paratyphoid) in domestic pigs [2, 3]. Although these other serovars continue to cause disease in southern and eastern Asia, Choleraesuis is rare in Europe today except in wild boar [5], and Typhisuis has been eradicated from European pigs. For our further phylogenetic analyses, we included two genomes of serovar Birkenhead as an outgroup because they were the closest genetic relatives of the Para C lineage (Figure 2).

Reliable inference of the evolutionary timescale and phylogeographic history of the Para C lineage depends on a broad temporal and spatial range of sources for the bacterial strains. However, EnteroBase only contained 100 *Salmonella* genomes from the Para C lineage, and they were of limited geographical and temporal diversity. We therefore combed the strain collection at the Institut Pasteur, Paris, and sequenced 119 additional Para-C-lineage genomes from diverse, historical sources. Our final dataset comprised 219 modern Para-C-lineage genomes, isolated between 1914 and 2015 from multiple continents (Table S1).

A maximum-likelihood tree of the core SNPs within the 219 genomes showed that they fell into well-defined sub-lineages within each serovar (Paratyphi C: PC-1, PC-2; Choleraesuis: CS Kunzendorf, CS *sensu stricto*, CS-3; Typhisuis: TS-1, TS-2) (Figure 2B). We also calculated a pan-genome, which was used to map the SK152 metagenomic reads after initial processing and de-duplication. Mapping identified 1,030,108 unique *Salmonella* reads in teeth (0.05%–0.18% of all reads) or the femur (0.01%), but not in dental calculus (Table 1). All of these reads were specific to Paratyphi C and covered 98.4% of a reference Paratyphi C genome (RKS4594) with a mean read depth of 7.3-fold (Table 1). To avoid spurious SNP calls associated with DNA damage, we only called SNPs in the Ragna genome that were covered by at least two reads, resulting in 95% coverage of RKS4594 (Figure S2).

**Table 1. Reads Specific to *S. enterica* within Metagenomic Sequences of Samples from SK152**

| | | | | | Reads specific for *S. enterica* (Ragna) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Source | No. of libraries | Total unique reads | Total human reads | Total non-human reads | % of all reads | % duplicates | No. of unique reads | Mean read length (bp) | Genome coverage |
| Upper 3rd left molar root and pulp | 2 | 237,735,419 | 58,068,866 | 179,666,553 | 0.050 | 77 | 26,853 | 56 | 0.29 |
| Upper 2nd right molar dentine/cementum | 4 | 1,077,156,946 | 127,986,372 | 949,170,574 | 0.183 | 53 | 920,267 | 35 | 6.39 |
| Upper 2nd right molar pulp | 1 | 119,308,674 | 26,725,529 | 92,583,145 | 0.088 | 26 | 77,928 | 43 | 0.60 |
| Femoral long bone | 1 | 73,372,819 | 12,765,129 | 60,607,690 | 0.013 | 49 | 5,060 | 41 | 0.04 |
| Dental calculus (multiple teeth) | 1 | 235,375,745 | 20,737,655 | 214,638,090 | 0.000 | N/A | N/A | N/A | N/A |
| Total: | 9 | 1,742,949,603 | 246,283,551 | 1,496,666,052 | 0.126 | 53 | 1,030,108 | 36 | 7.32 |

See also Figure S1 for archaeological information and dating estimates for the burial of SK152. N/A: Not applicable

Our data demonstrate that Paratyphi C bacteria caused human infections in Norway 800 years ago, and their presence in both teeth and bones suggests that SK152 died of septicemia associated with enteric fever. Paratyphi C aDNA from 1,545 CE has also been recently described from mass graves in Mexico [19], consistent with a continuous history of systemic human disease associated with this pathogen.

## Pan-genomic Stability

The selective pressures associated with local ecological interactions are thought to cause variation of gene content in microbes [20]. We therefore anticipated that 800 years of evolution would have resulted in dramatic differences in gene content between Ragna and modern Paratyphi C genomes, and we expected even greater differences between Paratyphi C and the other clades of the Para C lineage. Surprisingly, 78% of the 4,388 ± 99 genes (total length 4.8 ± 0.08 Mb) in a Para-C-lineage genome were intact core genes, and only 604 core SNPs distinguished Ragna from the MRCA of modern Paratyphi C (Figure 3). Some core genes are universally present in the Para C lineage plus Birkenhead even though they belong to mobile genetic elements that are variably present in other *Salmonella*, e.g., the pathogenicity islands SPI-1 to SPI-6, SPI-9, and SPI-11 to SPI-14 (Figure 2B). Similarly, the virulence plasmid was present throughout the Para C lineage except for Typhisuis sub-lineage TS-2. A further constant feature of the Para C lineage was the absence of genes encoding typhoid toxin, which is thought to trigger enteric fever by serovars Typhi and Paratyphi A [21].

Other studies have indicated that microbial host adaptation is accompanied by the accumulation of pseudogenes [22], driven by the streamlining of genes that are no longer necessary for the infection of multiple hosts [23], or by rewiring of transcriptional regulation [24]. The 2,964 representative *Salmonella* genomes contained a median of 40–60 pseudogenes. The numbers of pseudogenes were unexceptional for the most recent common ancestors (MRCAs) of the Para C lineage (25 pseudogenes) or of Paratyphi C, Choleraesuis plus Typhisuis (69 pseudogenes) (Figure S3C), suggesting that neither of these MRCAs was adapted to any particular host. However, the MRCAs of the individual serovars may well mark the beginnings of host adaptation because they were associated with higher numbers of pseudogenes (Choleraesuis: 95; Paratyphi C: 116; Typhisuis: 181). These pseudogenes may simply represent functions that are not required for infection of their individual hosts or may even have contributed to host specificity.

We also attempted to identify mobile genetic elements in the accessory genome of the Para C lineage that could account for their differential host specificities. The 3,901 accessory genes clustered together within 227 GIs (genomic islands and other mobile elements), including 37 plasmids, 32 prophages, 16 IMEs (integrative and mobilizable elements), SPI-5 to SPI-7, and two ICEs (integrative and conjugative elements) (Table S2). Parts or all of these GIs were acquired or lost on 311 independent occasions. However, at least 60% of gains or losses are unlikely to be important for host specificity because they were restricted to a single genome (Table S2), and most gains or losses were very recent (Figure S3B). Most of the other gains or losses are also unlikely to represent successful evolutionary changes in virulence or host specificity because, as in other *Salmonella* [25–27], they were restricted to individual sub-lineages, and sister sub-lineages differing in the possession of those genes are also prevalent in invasive disease (Figure 2B). For example, Paratyphi C is more virulent for mice after lysogenization upstream of *pgtE* by a P22-like prophage, SCP-P1 [28] (here GI076), whose gene product prevents opsonization [29]. However, GI076 is absent from half of the Paratyphi C genomes, including Ragna (Figure 2B). Similarly, the *fimH102* allele of a type 1 fimbrial adhesin facilitates specific adhesion to porcine cells by serovar Choleraesuis [30], but *fimH* is totally lacking in CS *sensu stricto*. Indeed, none of the inferred virulence factors and GIs seemed likely to be consistently related to differential virulence or host specificity (Table S3), with the notable exceptions of SPI-7 and SPI-6.

SPI-7 (GI107) is a pathogenicity island which encodes the Vi capsular polysaccharide in serovars Typhi, Paratyphi C (including Ragna), and Dublin. Vi might promote enteric fever in humans because it prevents the opsonization and clearance that is triggered by binding of the C3 component of complement to lipopolysaccharide [31]. The presence of SPI-7 in all Paratyphi C, and its absence from all Typhisuis and Choleraesuis, suggests

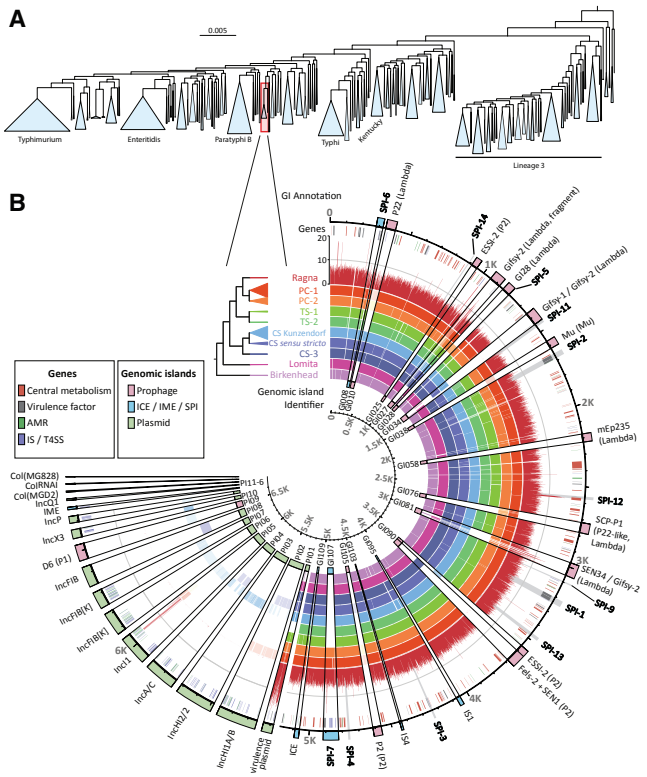


**Figure 2. Genomic Phylogenies of *Salmonella enterica* and the Para C Lineage**

(A) Maximum-likelihood phylogeny of 2,964 representative genomes of *S. enterica* subspecies *enterica*. Each genome is a representative of one ribosomal multi-locus sequence typing (rMLST) sequence type. Blue triangles indicate common lineages containing numerous genomes, including the Para C lineage. Several lineages are associated with common serovars, as indicated by horizontal labels. Red rectangle: Para C lineage plus the outgroup Birkenhead.

(B) Pan-genomic contents for 6,665 pan-genes in 222 genomes of the Para C lineage, including Ragna, plus Birkenhead, with one stroke per gene. Circles (inner to outer); Circle 1: sixteen major, variably present chromosomal genomic islands (GI008–GI109) followed by sixteen cytoplasmic plasmids, circular phages plus one IME (PI01–PI16), color-coded as in the genomic islands key. Circles 2–10: the frequency of presence or absence of each gene per sub-lineage within the phylogram is indicated by color opacity. Circle 11: coverage of aDNA reads per gene within Ragna (scale 0–20 reads at 12:00). Circle 12: genes color-coded as in the genes key. Circle 13: traditional designations for GIs, PIs, and other variably present genomic elements (Table S3). Gray wedges: SPIs.

See also Table S1 for summary statistics on the sources and dates of collection of the bacterial strains and Table S5 for the metadata for each individual genome.

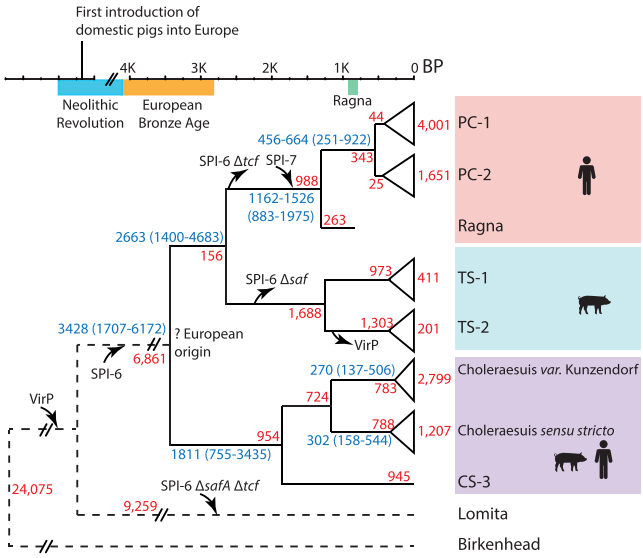


**Figure 3. Cartoon of the Evolutionary History of the Para C Lineage on a Time Frame for Human History in Europe**

BP: before present. The tree is annotated with the acquisitions of the virulence plasmid (VirP), SPI-6 and SPI-7 as annotated by inward arrows, and deletions of parts of SPI-6 as annotated by outward arrows. Numbers in blue indicate date estimates and their 95% credible intervals (parentheses) according to a Bayesian phylogenetic approach. Red numbers indicate numbers of substitution events for non-recombinant core SNPs except for Lomita and Birkenhead, where they indicate total core SNPs. The host specificities of the individual serovars are indicated by cartoons at the right.

See also Figure S3 for the timing of individual changes in pseudogenes and gene gain/loss, Figure S4 for the locations of variable genes by lineage within SPI-6, Table S3 for the properties of individual genomic islands, and Table S4 for detailed estimates of tMRCAs. See also Table S2.

that SPI-7 was acquired prior to the expansion of serovar Paratyphi C and also suggests an association with its human specificity. A 5-gene deletion within SPI-7 is present throughout sub-lineage PC-1, but this does not affect the production of Vi polysaccharide [3].

SPI-6 (GI008) is present throughout the Para C lineage. It encodes a type-6 secretion system (T6SS), as well as *Salmonella* atypical fimbriae (*saf*) and Typhi colonization factor (*tcf*) (Figure S4). T6SS systems encode intracellular, inverted bacteriophage-tail-like structures that can inject lethal effectors (TaeX) into neighboring eukaryotic and bacterial cells [32]. SPI-6 contributes to the gastrointestinal colonization and pathogenesis of mice [33] and chickens [34] by serovar Typhimurium. Expression of *tcf* resulted in specific adhesion of *Escherichia coli* to human epithelial cells [35], and *saf* mutations reduced gastrointestinal colonization of pigs by serovar Choleraesuis [36]. The *tcf* and *saf* genes are variably present within SPI-6 in the multiple serovars within the Para C lineage (Figure S4) and might therefore also account for their differential host specificity. A parsimonious interpretation of the origins of this diversity is that an intact SPI-6 was initially acquired by horizontal transfer after Lomita branched off, followed by successive, internal deletions prior to the MRCAs of Paratyphi C and Typhisuis. Alternatively, multiple SPI-6 variants might each represent an independent horizontal transfer event.

## Evolutionary Timing

These observations immediately raised the question of evolutionary timescales, including the ages of individual serovars. We confirmed the existence of significant temporal signal according to root-to-tip distances [37] and date randomization tests [38] with the non-recombinant SNPs from the Para-C-lineage genomes from 10 independent samples of Paratyphi C (both including and excluding Ragna) and also from the Para C lineage (without Lomita, which lacks a collection date). We then dated

key stages in the evolution of the Para C lineage by a Bayesian phylogenetic approach (Figure 3). The optimal model for Paratyphi C (strict clock) yielded a median substitution rate of $7.9 \times 10^{-8}$ substitutions/site/year (95% credible interval: $5.3 \times 10^{-8}$-$1.1 \times 10^{-7}$), slightly lower than the median rate estimated according to the optimal model (relaxed) for the Para C lineage ($1.5 \times 10^{-7}$; 95% CI: $6.9 \times 10^{-8} - 2.5 \times 10^{-7}$) (Table S4). The age of the crown group encompassing modern isolates of Paratyphi C was dated at 456–664 BP (95% CI: 251–922) and its split from Ragna at 1,162–1,526 (95% CI: 883-1,975) (Table S4). Thus, the addition of the Ragna genome sheds light on about 800 additional years of evolutionary history of Paratyphi C.

Our estimate of the timescale allowed us to investigate the rates of gain or loss events and pseudogene formation within Paratyphi C relative to the accumulation of SNPs (Figure S3). The results showed that pseudogenes accumulated slowly and continuously (3–5 per 100 SNPs) since the tMRCA until the last 200 years, when their rate accelerated to 7 pseudogenes/100 SNPs (Figure S3A). The rate of gene gain or loss was also low (1.3 genes/100 SNPs) until 400 years ago, increased to 15–20 genes/100 SNPs and increased even further to 49 genes/100 SNPs in the last 50 years before laboratory cultivation (Figure S3B). Similar results were obtained for the other sub-lineages within the Para C lineage (Figures S3C and S3D). Recent gene gain or loss has also been noted in *S. enterica* serovars Paratyphi A [25] and Agona [26] and was attributed to frequent acquisitions of selfish DNA, which were also subsequently rapidly lost. Our results confirm that genomic islands are indeed highly variable in modern isolates. However, they clearly show that they were lost or gained much less often in previous millennia, which were marked by relative pan-genomic stability.

The age of the stem lineage of Paratyphi C stretches back to its differentiation from serovar Typhisuis, about 2,663 years (95% CI: 1,440–4,683), during which time ~1,350 SNPs were accumulated prior to the MRCAs of the PC-1 and PC-2 clades (Figure 3). In turn, the estimated tMRCA for Paratyphi C, Typhisuis plus Choleraesuis was approximately 3,428 years ago (95% CI: 1,707–6,172) and that ancestor evolved in Europe according to three independent Bayesian and maximum-likelihood methods. A European ancestry is consistent with the existence of Paratyphi C and enteric fever in northern Norway 800 years ago and also with enteric fever caused by Paratyphi C bacteria in Mexico in 1,545 CE, which was inferred to have been introduced by Europeans [19]. We also note that serovars Choleraesuis and Typhisuis infect swine and that their tMRCAs overlap with the Neolithic domestication of pigs from wild boar in Europe [39] (Figure 3). It is therefore possible that Paratyphi C represents the host specialization to humans of a zoonotic pathogen of domesticated animals. Alternatively, Typhisuis may have specialized from a generalist life style to a host specificity for swine.

Our identification of the Ragna genome from a young woman who died in 1,200 CE provides insights on the genomic contents of the stem lineage of serovar Paratyphi C. The Ragna genome also demonstrates that salmonellosis was a deadly invasive disease of humans for centuries before its first recognition by physicians. Our analyses show that reconstructing the long-term evolutionary history of bacterial pathogens benefits dramatically from comparisons of metagenomic data from ancient samples with population genetic data from present-day bacteria. The close relationship between clades of the Para C lineage that differ in host specificity triggered intriguing speculations about historical host jumps during the Neolithic period between humans and their domesticated animals. Our results also indicate that both the core and accessory genome of bacterial pathogens can be remarkably stable over millennia and that much of the dramatic variation between extant genomes represents transient genetic fluctuation, whose evolutionary relevance to ecological fitness is uncertain.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - SK152
- METHOD DETAILS
  - Dating of SK152
  - Metagenomic sequencing of samples from SK152
  - Genomic assemblies from metagenomic sequences of samples from SK152
  - Genotyping by EnteroBase
  - Sources of genomes from the Para C Lineage
  - *Salmonella* short reads from metagenomic aDNA
  - Date estimates for Paratyphi C and related serovars
  - Inferring ancestral geographic sources
  - Reconstruction of the pan genome of the Para C lineage
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

## SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and five tables and can be found with this article online at https://doi.org/10.1016/j.cub.2018.05.058.

## ACKNOWLEDGMENTS

EnteroBase (BBSRC BB/L020319/1) was developed by N.-F.A., M.J.S., and Z.Z. (equal contributions) under guidance by M.A. Additional grant support was from the Wellcome Trust (202792/Z/16/Z), Medical Research Council (MR/M50161X/1), Grundforksningfonden (DNRF 94) and Lundbeckfonden (R52-5062), the French government (ANR-10-LABX-62-IBEID), the Institut Pasteur, Santé Publique France, and Fondation Le Roch-Les Mousquertaires. We thank A. Murat Eren for assistance with implementation of Anvi'o in EnteroBase, and we thank Siavash Mirarab and Tandy Warnow for advice on Astrid and running Astral-II. We thank Birgitte Skar for permission to sample the skeletal material and the Danish National High-Throughput DNA Sequencing Centre for technical assistance with data generation. We thank Monica H. Green, Timothy P. Newfield, and Francois Balloux for helpful comments on the manuscript.

## AUTHOR CONTRIBUTIONS

M.A., S.Y.W.H., F.-X.W., M.T.P.G., and Z.Z. designed experiments. S.N., H.K.S., S.S.H., A.C., N.R.T., F.-X.W., and M.T.P.G. provided samples for analysis. I.L., A.T.-D., M.J.S., G.L., and A.K.F. performed laboratory work. Z.Z., S.D., S.Y.W.H., N.-F.A., S.C., and C.Q. performed analyses. M.A., Z.Z., I.L., N.-F.A., C.Q., S.Y.W.H., and M.T.P.G. wrote the paper.

## DECLARATIONS OF INTERESTS

The authors declare no competing interests.

Received: November 3, 2017
Revised: February 9, 2018
Accepted: May 18, 2018
Published: July 19, 2018

## REFERENCES

1. Jacobs, M.R., Koornhof, H.J., Crisp, S.I., Palmhert, H.L., and Fitzstephens, A. (1978). Enteric fever caused by *Salmonella* paratyphi C in South and South West Africa. S. Afr. Med. J. *54*, 434–438.
2. Uzzau, S., Brown, D.J., Wallis, T., Rubino, S., Leori, G., Bernard, S., Casadesús, J., Platt, D.J., and Olsen, J.E. (2000). Host adapted serotypes of *Salmonella enterica*. Epidemiol. Infect. *125*, 229–255.
3. Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M.G., Hale, J.L., Harbottle, H., Uesbeck, A., et al.; S. Enterica MLST Study Group (2012). Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. *8*, e1002776.
4. Hirschfeld, L. (1919). A new germ of paratyphoid. Lancet *193*, 296–297.
5. Methner, U., Heller, M., and Bocklisch, H. (2009). *Salmonella enterica* subspecies *enterica* serovar Choleraesuis in a wild boar population in Germany. Eur. J. Wildl. Res. *56*, 493–502.
6. Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., Vigne, J.D., Edwards, C.J., Schlumbaum, A., Dinu, A., et al. (2007). Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. Proc. Natl. Acad. Sci. USA *104*, 15276–15281.
7. McNeill, W.H. (1976). Plagues and peoples, First Edition (New York: Anchor).
8. Achtman, M. (2016). How old are bacterial pathogens? Proc. Biol. Sci. *283*, 1836.
9. Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jäger, G., Bos, K.I., Herbig, A., Economou, C., Benjak, A., Busso, P., et al. (2013). Genome-wide comparison of medieval and modern *Mycobacterium leprae*. Science *341*, 179–183.
10. Rasmussen, S., Allentoft, M.E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.G., Pedersen, A.G., Schubert, M., Van Dam, A., Kapel, C.M., et al. (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. Cell *163*, 571–582.
11. Forsåker, A.-L., and Göthberg, H. (1986). Stratigrafisk analyse, delfelt FJ, FN og FW (Trondheim: Riksantikvarens utgravningskontor for Trondheim).
12. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat. Methods *11*, 1144–1146.
13. Stuiver, M., and Braziunas, T.F. (1993). Modeling atmospheric $^{14}C$ influences and $^{14}C$ ages of marine samples to 10,000 BC. Radiocarbon *35*, 137–189.
14. Hamre, S.S., and Daux, V. (2016). Stable oxygen isotope evidence for mobility in medieval and post-medieval Trondheim, Norway. Journal of Archaeological Science: Reports *8*, 416–425.
15. Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R.G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. *45*, 6310–6320.
16. Downes, J., Munson, M.A., Spratt, D.A., Kononen, E., Tarkka, E., Jousimies-Somer, H., and Wade, W.G. (2001). Characterisation of *Eubacterium*-like strains isolated from oral infections. J. Med. Microbiol. *50*, 947–951.
17. Fegan, M. (2006). Plant pathogenic members of the genera *Acidovorax* and *Herbaspirillum*. In Plant-Associated Bacteria, S.S. Gnanamanickam, ed. (Springer Netherlands), pp. 671–702.
18. Alikhan, N.-F., Zhou, Z., Sergeant, M.J., and Achtman, M. (2018). A genomic overview of the population structure of *Salmonella*. PLoS Genet. *14*, e1007261.
19. Vågene, Å.J., Herbig, A., Campana, M.G., Robles Garcia, N.M., Warinner, C., Sabin, S., Spyrou, M.A., Andrades Valtueña, A., Huson, D., Tuross, N., et al. (2018). Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. Nat. Ecol. Evol. *2*, 520–528.
20. Cordero, O.X., and Polz, M.F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. Nat. Rev. Microbiol. *12*, 263–273.
21. Galán, J.E. (2016). Typhoid toxin provides a window into typhoid fever and the biology of *Salmonella* Typhi. Proc. Natl. Acad. Sci. USA *113*, 6338–6344.
22. McClelland, M., Sanderson, K.E., Clifton, S.W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., et al. (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. Nat. Genet. *36*, 1268–1274.
23. Bäumler, A., and Fang, F.C. (2013). Host specificity of bacterial pathogens. Cold Spring Harb. Perspect. Med. *3*, a010041.
24. Hottes, A.K., Freddolino, P.L., Khare, A., Donnell, Z.N., Liu, J.C., and Tavazoie, S. (2013). Bacterial adaptation through loss of function. PLoS Genet. *9*, e1003617.
25. Zhou, Z., McCann, A., Weill, F.X., Blin, C., Nair, S., Wain, J., Dougan, G., and Achtman, M. (2014). Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. Proc. Natl. Acad. Sci. USA *111*, 12199–12204.
26. Zhou, Z., McCann, A., Litrup, E., Murphy, R., Cormican, M., Fanning, S., Brown, D., Guttman, D.S., Brisse, S., and Achtman, M. (2013). Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. PLoS Genet. *9*, e1003471.
27. Langridge, G.C., Fookes, M., Connor, T.R., Feltwell, T., Feasey, N., Parsons, B.N., Seth-Smith, H.M., Barquist, L., Stedman, A., Humphrey, T., et al. (2015). Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc. Natl. Acad. Sci. USA *112*, 863–868.
28. Zou, Q.H., Li, Q.H., Zhu, H.Y., Feng, Y., Li, Y.G., Johnston, R.N., Liu, G.R., and Liu, S.L. (2010). SPC-P1: a pathogenicity-associated prophage of *Salmonella* paratyphi C. BMC Genomics *11*, 729.
29. Riva, R., Korhonen, T.K., and Meri, S. (2015). The outer membrane protease PgtE of *Salmonella enterica* interferes with the alternative complement pathway by cleaving factors B and H. Front. Microbiol. *6*, 63.
30. Yue, M., Han, X., De Masi, L., Zhu, C., Ma, X., Zhang, J., Wu, R., Schmieder, R., Kaushik, R.S., Fraser, G.P., et al. (2015). Allelic variation contributes to bacterial host specificity. Nat. Commun. *6*, 8754.
31. Wilson, R.P., Winter, S.E., Spees, A.M., Winter, M.G., Nishimori, J.H., Sanchez, J.F., Nuccio, S.P., Crawford, R.W., Tükel, Ç., and Bäumler, A.J. (2011). The Vi capsular polysaccharide prevents complement receptor 3-mediated clearance of *Salmonella enterica* serotype Typhi. Infect. Immun. *79*, 830–837.
32. Hachani, A., Wood, T.E., and Filloux, A. (2016). Type VI secretion and anti-host effectors. Curr. Opin. Microbiol. *29*, 81–93.
33. Sana, T.G., Flaugnatti, N., Lugo, K.A., Lam, L.H., Jacobson, A., Baylot, V., Durand, E., Journet, L., Cascales, E., and Monack, D.M. (2016). *Salmonella* Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. Proc. Natl. Acad. Sci. USA *113*, E5044–E5051.
34. Pezoa, D., Yang, H.J., Blondel, C.J., Santiviago, C.A., Andrews-Polymenis, H.L., and Contreras, I. (2013). The type VI secretion system encoded in SPI-6 plays a role in gastrointestinal colonization and systemic spread of *Salmonella enterica* serovar Typhimurium in the chicken. PLoS ONE *8*, e63917.

35. Leclerc, J.M., Quevillon, E.L., Houde, Y., Paranjape, K., Dozois, C.M., and Daigle, F. (2016). Regulation and production of Tcf, a cable-like fimbriae from *Salmonella enterica* serovar Typhi. Microbiology *162*, 777–788.

36. Carnell, S.C., Bowen, A., Morgan, E., Maskell, D.J., Wallis, T.S., and Stevens, M.P. (2007). Role in virulence and protective efficacy in pigs of *Salmonella enterica* serovar Typhimurium secreted components identified by signature-tagged mutagenesis. Microbiology *153*, 1940–1952.

37. Rambaut, A., Lam, T.T., Max, C.L., and Pybus, O.G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. *2*, vew007.

38. Ramsden, C., Melo, F.L., Figueiredo, L.M., Holmes, E.C., and Zanotto, P.M.; VGDN Consortium (2008). High rates of molecular evolution in hantaviruses. Mol. Biol. Evol. *25*, 1488–1492.

39. Krause-Kyora, B., Makarewicz, C., Evin, A., Flink, L.G., Dobney, K., Larson, G., Hartz, S., Schreiber, S., von Carnap-Bornheim, C., von Wurmb-Schwark, N., and Nebel, A. (2013). Use of domesticated pigs by Mesolithic hunter-gatherers in northwestern Europe. Nat. Commun. *4*, 2348.

40. Yoon, S.H., Park, Y.K., and Kim, J.F. (2015). PAIDB v2.0: exploration and analysis of pathogenicity and resistance islands. Nucleic Acids Res. *43*, D624–D630.

41. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. *34*, D32–D36.

42. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

43. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

44. Vachaspati, P., and Warnow, T. (2015). ASTRID: Accurate Species TRees from Internode Distances. BMC Genomics *16* (*Suppl 10*), S3.

45. Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

46. To, T.H., Jung, M., Lycett, S., and Gascuel, O. (2016). Fast dating using least-squares criteria and algorithms. Syst. Biol. *65*, 82–97.

47. Pagel, M. (2017). BayesTraits V3.0. *http://www.evolution.rdg.ac.uk/BayesTraitsV3/BayesTraitsV3.html*.

48. Ramsey, C.B., and Lee, S. (2013). Recent and planned developments of the program OxCal. In Proceedings of the 21st International Radiocarbon Conference 55, A.J.T. Jull, and C. Hatte, eds. (University of Arizona), pp. 720–730.

49. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

50. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics *31*, 1674–1676.

51. Applegate, D.L., Bixby, R.E., Chvátal, V., and Cook, W.J. (2006). The Traveling Salesman Problem - A Computational Study (New Jersey: Princeton University Press).

52. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

53. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

54. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

55. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

56. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

57. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

58. Kay, G.L., Sergeant, M.J., Zhou, Z., Chan, J.Z., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M., Loman, N.J., et al. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat. Commun. *6*, 6717.

59. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

60. Bushnell, B. (2016). BBMap short read aligner. https://sourceforge.net/projects/bbmap.

61. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

62. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

63. Hamada, M., Wijaya, E., Frith, M.C., and Asai, K. (2011). Probabilistic alignments with quality scores: an application to short-read mapping toward accurate SNP/indel detection. Bioinformatics *27*, 3085–3092.

64. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

65. Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18.

66. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. *33*, 1635–1638.

67. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.

68. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

69. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

70. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. *44* (W1), W16–21.

71. Carattoli, A., Zankari, E., García-Fernández, A., Voldby Larsen, M., Lund, O., Villa, L., Møller Aarestrup, F., and Hasman, H. (2014). *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob. Agents Chemother. *58*, 3895–3903.

72. Guglielmini, J., Néron, B., Abby, S.S., Garcillán-Barcia, M.P., de la Cruz, F., and Rocha, E.P. (2014). Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. *42*, 5715–5727.

73. Eren, A.M., Esen, O.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319.

74. Joshi, N.A., and Fass, J.N. (2016). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). *https://github.com/najoshi/sickle*.

75. Yu, Y., Harris, A.J., Blair, C., and He, X. (2015). RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Mol. Phylogenet. Evol. *87*, 46–49.

76. McKinley, J.I. (2004). Compiling a skeletal inventory: disarticulated and co-mingled remains. In Guidelines to the standards for recording human remains, M. Brickley, and J.I. McKinley, eds. (Southampton: CIfA Papers), pp. 14–17.

77. Cox, M. (2000). Ageing adults from the skeleton. In Human osteology: in archaeology and forensic science, M. Cox, and S. Mays, eds. (London: Cambridge University Press), pp. 61–82.

78. Mays, S., and Cox, M. (2000). Sex determination in skeletal remains. In Human osteology: in archaeology and forensic science, M. Cox, and S. Mays, eds. (London: Cambridge University Press), pp. 117–130.

79. Skoglund, P., Storå, J., Götherström, A., and Jakobsson, M. (2013). Accurate sex identification of ancient human remains using DNA shotgun sequencing. J. Archaeol. Sci. *40*, 4477–4482.

80. Daux, V., Lécuyer, C., Héran, M.A., Amiot, R., Simon, L., Fourel, F., Martineau, F., Lynnerup, N., Reychler, H., and Escarguel, G. (2008). Oxygen isotope fractionation between human phosphate and water revisited. J. Hum. Evol. *55*, 1138–1147.

81. Chenery, C.A., Pashley, V., Lamb, A.L., Sloane, H.J., and Evans, J.A. (2012). The oxygen isotope relationship between the phosphate and structural carbonate fractions of human bioapatite. Rapid Commun. Mass Spectrom. *26*, 309–319.

82. Christophersen, A., Jondell, E., Marstein, O., Nordeide, S.W., and Reed, I. (1988). Utgravning, kronologi og bebyggelsesutvikling (Trondheim: Riksantikvarens utgravningskontor for Trondheim).

83. Christophersen, A., and Nordeide, S.W. (1994). Kaupangen ved Nidelva: 1000 a°rs byhistorie belyst gjennom de arkeologiske undersøkelsene pa° Folkebibliotekstomten i Trondheim 1973-1985 (Oslo: Riksantikvaren).

84. Ramsey, C., Higham, T., and Leach, P. (2016). Towards high-precision AMS: progress and limitations. Radiocarbon *46*, 17–24.

85. Ramsey, C.B., Higham, T.F., Owen, D.C., Pike, A.W.G., and Hedges, R.E.M. (2016). Radiocarbon dates from the Oxford AMS system: archaeometry datelist 31. Archaeometry *44*, 1–149.

86. Reimer, P.J., Bard, E., Bayliss, A., Beck, J.W., Blackwell, P.G., Ramsey, C.B., Buck, C.E., Cheng, H., Edwards, R.L., Friedrich, M., et al. (2013). IntCal13 and Marine13 radiocarbon age calibration curves 0-50,000 years cal BP. Radiocarbon *55*, 1869–1887.

87. Rohland, N., and Hofreiter, M. (2007). Ancient DNA extraction from bones and teeth. Nat. Protoc. *2*, 1756–1762.

88. Allentoft, M.E., Sikora, M., Sjögren, K.G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature *522*, 167–172.

89. Campbell, M.A., Van Leuven, J.T., Meister, R.C., Carey, K.M., Simon, C., and McCutcheon, J.P. (2015). Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont *Hodgkinia*. Proc. Natl. Acad. Sci. USA *112*, 10192–10199.

90. Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283–1287.

91. Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, A.J., and Maiden, M.C. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology *158*, 1005–1015.

92. Cadillo-Quiroz, H., Yavitt, J.B., and Zinder, S.H. (2009). *Methanosphaerula palustris* gen. nov., sp. nov., a hydrogenotrophic methanogen isolated from a minerotrophic fen peatland. Int. J. Syst. Evol. Microbiol. *59*, 928–935.

93. Rey, F.E., Gonzalez, M.D., Cheng, J., Wu, M., Ahern, P.P., and Gordon, J.I. (2013). Metabolic niche of a prominent sulfate-reducing human gut bacterium. Proc. Natl. Acad. Sci. USA *110*, 13582–13587.

94. Del Fabbro, C., Scalabrin, S., Morgante, M., and Giorgi, F.M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS ONE *8*, e85024.

95. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

96. Achtman, M., Zhou, Z., and Didelot, X. (2015). Formal comment to Pettengill: The time to Most Recent Common Ancestor does not (usually) approximate the date of divergence. PLoS ONE *10*, e0134435.

97. Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics *31*, i44–i52.

98. Baker, K.S., Burnett, E., McGregor, H., Deheer-Graham, A., Boinett, C., Langridge, G.C., Wailan, A.M., Cain, A.K., Thomson, N.R., Russell, J.E., and Parkhill, J. (2015). The Murray collection of pre-antibiotic era *Enterobacteriacae*: a unique research resource. Genome Med. *7*, 97.

99. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical assessment of metagenome interpretation - a benchmark of computational metagenomics software. Nat. Methods *14*, 1063–1071.

100. Leaché, A.D., Banbury, B.L., Felsenstein, J., de Oca, A.N., and Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for Inferring SNP phylogenies. Syst. Biol. *64*, 1032–1047.

101. Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proc. Biol. Sci. *255*, 37–45.

102. Didelot, X., and Wilson, D.J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. *11*, e1004041.

103. Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., and Harris, S.R. (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. *43*, e15.

104. Minin, V.N., Bloomquist, E.W., and Suchard, M.A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. *25*, 1459–1471.

105. Stadler, T. (2010). Sampling-through-time in birth-death trees. J. Theor. Biol. *267*, 396–404.

106. Xie, W., Lewis, P.O., Fan, Y., Kuo, L., and Chen, M.H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. *60*, 150–160.

107. Duchêne, S., Duchêne, D., Holmes, E.C., and Ho, S.Y. (2015). The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. Mol. Biol. Evol. *32*, 1895–1906.

108. Pagel, M. (1999). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. *48*, 612–622.

109. Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. *53*, 673–684.

110. Li, Q., Hu, Y., Xu, L., Xie, X., Tao, M., and Jiao, X. (2014). Complete genome sequence of *Salmonella enterica* serovar Choleraesuis vaccine strain C500 Attenuated by chemical mutation. Genome Announc. *2*, e1022-14.

111. Chiu, C.H., Tang, P., Chu, C., Hu, S., Bao, Q., Yu, J., Chou, Y.Y., Wang, H.S., and Lee, Y.S. (2005). The genome sequence of *Salmonella enterica* serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen. Nucleic Acids Res. *33*, 1690–1698.

112. Liu, W.Q., Feng, Y., Wang, Y., Zou, Q.H., Chen, F., Guo, J.T., Peng, Y.H., Jin, Y., Li, Y.G., Hu, S.N., et al. (2009). *Salmonella paratyphi* C: genetic divergence from *Salmonella choleraesuis* and pathogenic convergence with *Salmonella typhi*. PLoS ONE *4*, e4510.

113. Casjens, S.R., and Grose, J.H. (2016). Contributions of P2- and P22-like prophages to understanding the enormous diversity and abundance of tailed bacteriophages. Virology *496*, 255–276.

114. Blondel, C.J., Jiménez, J.C., Contreras, I., and Santiviago, C.A. (2009). Comparative genomic analysis uncovers 3 novel loci encoding type six

secretion systems differentially distributed in *Salmonella* serotypes. BMC Genomics *10*, 354.

115. Vernikos, G.S., and Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. Bioinformatics *22*, 2196–2203.

116. Elder, J.R., Chiok, K.L., Paul, N.C., Haldorson, G., Guard, J., and Shah, D.H. (2016). The *Salmonella* pathogenicity island 13 contributes to pathogenesis in streptomycin pre-treated mice but not in day-old chickens. Gut Pathog. *8*, 16.

117. Shah, D.H., Lee, M.J., Park, J.H., Lee, J.H., Eo, S.K., Kwon, J.T., and Chae, J.S. (2005). Identification of *Salmonella* gallinarum virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis. Microbiology *151*, 3957–3968.

118. Fuentes, J.A., Villagra, N., Castillo-Ruiz, M., and Mora, G.C. (2008). The *Salmonella* Typhi *hlyE* gene plays a role in invasion of cultured epithelial cells and its functional transfer to *S.* Typhimurium promotes deep organ infection in mice. Res. Microbiol. *159*, 279–287.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Bacterial and Virus Strains | | |
| Table S5. Bacterial Strains and genomes.xlsx | This paper | N/A |
| Deposited Data | | |
| Ancestral reconstruction 1<br>Maximum likelihood phylogeny and geographical origins derived from 49,610 non-recombinant, non-repetitive core SNPs within the Para C lineage plus Birkenhead | This paper | http://wrap.warwick.ac.uk/101809 |
| Ancestral reconstruction 2<br>Geographic inference of the sources of clades within the Para C lineage | This paper | http://wrap.warwick.ac.uk/101810 |
| Concoct 1<br>Single core gene (SCG) frequencies in the 76 CONCOCT clusters generated after binning contigs | This paper | http://wrap.warwick.ac.uk/101811 |
| Concoct 2<br>Statistics after clustering all metagenomes with MEGAHIT | This paper | http://wrap.warwick.ac.uk/101812 |
| Concoct 3<br>Taxonomic assignments for 11 metagenome assembled genomes (MAGs) | This paper | http://wrap.warwick.ac.uk/101813 |
| Concoct 4<br>Single-stranded deamination rates for Ragna, human DNA and 11 MAGs | This paper | http://wrap.warwick.ac.uk/101814 |
| Concoct 5<br>Genomic coverage of 11 MAGs by source and sequencing library | This paper | http://wrap.warwick.ac.uk/101815 |
| Concoct 6<br>Bacterial taxa found by Kraken in SK152 metagenomes at $\geq$ 0.02% frequency. | This paper | http://wrap.warwick.ac.uk/101816 |
| Concoct 7<br>Eleven MAGs Contigs. | This paper | http://wrap.warwick.ac.uk/101817 |
| Concoct 8<br>Alignments and phylogenies of SCGs across the Tree of Life. | This paper | http://wrap.warwick.ac.uk/101818 |
| Date estimation 1<br>Date randomization tests for the temporal signal in multiple replicate datasets that were analyzed by BEAST | This paper | http://wrap.warwick.ac.uk/101819 |
| Date estimation 2<br>Clock-like behavior of root to tip distances for Ragna, Tepos-14 and Tepos-35 plus modern genomes from the Para C lineage or Paratyphi C. | This paper | http://wrap.warwick.ac.uk/101820 |
| Date estimation 3<br>Comparison of MRCA dating estimates by BEAST using ancient plus modern, and only modern strains | This paper | http://wrap.warwick.ac.uk/101821 |
| Date estimation 4<br>Raw Maximum Clade Credibility trees from BEAST inferences. | This paper | http://wrap.warwick.ac.uk/101822 |
| ParaC genome SNPs 1<br>Numbers of mutational and recombinational SNPs per clade within the Para C lineage plus Birkenhead | This paper | http://wrap.warwick.ac.uk/101831 |
| ParaC genome SNPs 2<br>SNPs in the core genome of the Para C lineage plus Birkenhead. | This paper | http://wrap.warwick.ac.uk/101832 |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| ParaC genome SNPs 3<br>Maximum likelihood phylogeny and inferred geographical origins based on 49,610 non-recombinant, non-repetitive core SNPs from the Para C lineage plus Birkenhead. | This paper | http://wrap.warwick.ac.uk/101833 |
| ParaC genome SNPs 4<br>Venn diagram of substitutions scored as recombinational or mutational by RecHMM, ClonalFrameML and Gubbins. | This paper | http://wrap.warwick.ac.uk/101834 |
| ParaC Pan-genome 1<br>Gain and loss of genomic islands shown on a maximum-likelihood radial phylogeny derived from 49,610 non-recombinant, non-repetitive core SNPs within the Para C lineage plus Birkenhead | This paper | http://wrap.warwick.ac.uk/101823 |
| ParaC Pan-genome 2<br>Summary statistics for pan-genomic contents of Ragna plus 219 modern genomes of the Para C lineage | This paper | http://wrap.warwick.ac.uk/101824 |
| ParaC Pan-genome 3<br>Gene gain/loss events and numbers of pseudogenes by sub-lineage | This paper | http://wrap.warwick.ac.uk/101825 |
| ParaC Pan-genome 4<br>The pan-genome of the Para C lineage based on 6,665 single copy genes from 220 genomes. | This paper | http://wrap.warwick.ac.uk/101826 |
| ParaC Pan-genome 5<br>Sequences, original gene names and annotations of the reference sequences for 6,665 pan genes in the Para C lineage. | This paper | http://wrap.warwick.ac.uk/101827 |
| ParaC Pan-genome 6<br>Pseudogenes associated with sub-lineages and clades of the Para C lineage. | This paper | http://wrap.warwick.ac.uk/101828 |
| ParaC Pan-genome 7<br>Genomic islands and plasmids. | This paper | http://wrap.warwick.ac.uk/101829 |
| ParaC Pan-genome 8<br>Data and script for extracting rate of gene gain/loss and pseudogenes from the branches of an ML tree. See readme.txt for instructions. | This paper | http://wrap.warwick.ac.uk/101830 |
| *Salmonella* supertree 1<br>Metadata for 2,964 genomes that are representative of the genomic diversity of *Salmonella enterica* subsp. I. | This paper | http://wrap.warwick.ac.uk/101835 |
| *Salmonella* supertree 2<br>A species tree (ASTRID) based on 3,002 core gene trees from 2,964 representative genomes from *S. enterica* subsp. I. | This paper | http://wrap.warwick.ac.uk/101836 |
| *Salmonella* supertree 3<br>A RAxML maximum likelihood phylogeny based on a 2.8 Mbp concatenate of 3,002 core genes from 2,964 representative genomes from *S. enterica* subsp. I | This paper | http://wrap.warwick.ac.uk/101837 |
| SK152 reads: Non-human metagenomic data (SK152) | This paper | https://sid.erda.dk/wsgi-bin/ls.py?share_id=E56xgi8CEl |
| Workflow<br>Workflow to reconstruct the Ragna genome | This paper | http://wrap.warwick.ac.uk/101838 |
| Software and Algorithms | | |
| TemporalFreq.py and data: A script for extracting rate of gene gain/loss and pseudogenes from the branches of an ML tree to generate data for Figure S3A,B | This paper | https://github.com/zheminzhou/TemporalCurve |
| PAIDB | [40] | http://www.paidb.re.kr/ |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| ISfinder | [41] | https://www-is.biotoul.fr/ |
| Concoct | [12] | https://github.com/BinPro/CONCOCT |
| Kraken | [42] | https://ccb.jhu.edu/software/kraken/ |
| RAxML v8.2.4 | [43] | https://sco.h-its.org/exelixis/software.html |
| Astrid | [44] | https://github.com/pranjalv123/ASTRID/ |
| Beast 1.8.3 | [45] | https://github.com/beast-dev/beast-mcmc/releases |
| LSD | [46] | http://www.atgc-montpellier.fr/LSD |
| BayesTraits | [47] | http://www.evolution.rdg.ac.uk/BayesTraits.html |
| OxCal 4.2 | [48] | http://intcal.qub.ac.uk/intcal13/ |
| BWA | [49] | http://bio-bwa.sourceforge.net/ |
| Megahit | [50] | https://github.com/voutcn/megahit |
| Concorde | [51] | www.math.uwaterloo.ca/tsp/concorde.html |
| Prodigal | [52] | https://github.com/hyattpd/Prodigal |
| Mafft | [53] | https://mafft.cbrc.jp/alignment/software/ |
| Trimal | [54] | http://trimal.cgenomics.org/ |
| FastTree | [55] | www.microbesonline.org/fasttree/ |
| MapDamage 2.0 | [56] | https://ginolhac.github.io/mapDamage/ |
| EnteroBase | [18] | https://enterobase.warwick.ac.uk/ |
| Spades 3.5 | [57] | http://spades.bioinf.spbau.ru/release3.5.0/ |
| MGPlacer | [58] | https://sourceforge.net/projects/mgplacer/ |
| BLAST | [59] | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ |
| BBmerge & BBduk2 in BBmap | [60] | https://sourceforge.net/projects/bbmap/ |
| Bowtie2 | [61] | http://bowtie-bio.sourceforge.net/bowtie2/ |
| SAMtools/BCFtools 1.2 | [62] | http://www.htslib.org/ |
| Last | [63] | http://last.cbrc.jp/ |
| TRF | [64] | https://tandem.bu.edu/trf/trf.html |
| PILER-CR | [65] | https://www.drive5.com/pilercr/ |
| RecHMM | [25] | https://github.com/zheminzhou/EToKi |
| Date randomization test | [38] | N/A |
| TempEst | [37] | http://tree.bio.ed.ac.uk/software/tempest/ |
| ETE3 Python package | [66] | http://etetoolkit.org/ |
| PROKKA | [67] | https://github.com/tseemann/prokka |
| UClust | [68] | https://www.drive5.com/usearch |
| APE package of R | [69] | http://ape-package.ird.fr/ |
| PHASTER | [70] | http://phaster.ca/ |
| PlasmidFinder | [71] | https://cge.cbs.dtu.dk/services/PlasmidFinder |
| CONJscan-T4SSscan | [72] | https://research.pasteur.fr/en/software/conjscan-t4ssscan/ |
| Anvi'o | [73] | http://merenlab.org/software/anvio/ |
| Sickle | [74] | https://github.com/najoshi/sickle |
| RASP | [75] | http://mnh.scu.edu.cn/soft/blog/RASP |
| Other | | |
| Pan-Genome: Interactive Anvi'o Plot for the pan genome of the Para C lineage. | This paper | https://enterobase.warwick.ac.uk/anvio/public/zhemin/ParaC_pangenome |
| Para C lineage: Interactive 220 genomes in the Para C lineage plus 2 Birkenhead genomes | This paper | http://enterobase.warwick.ac.uk/species/senterica/search_strains?query=workspace:3246 |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Ragna: Ragna genome | This paper | http://enterobase.warwick.ac.uk/species/senterica/search_strains?query=workspace:3246 |
| rST representatives: Interactive workspace for 2,964 genomes of one representative per ribosomal MLST ST in *Salmonella enterica* subsp. I | This paper | http://enterobase.warwick.ac.uk/species/senterica/search_strains?query=workspace:3247 |
| SPI-6: Interactive Anvi'o Plot for the SPI-6 genomic island in the 221 genomes | This paper | https://enterobase.warwick.ac.uk/anvio/public/zhemin/ParaC_SPI6 |

## CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Mark Achtman (m.achtman@warwick.ac.uk).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

Permission to undertake genetic analysis on SK152 was granted to MTP Gilbert by the Norwegian Research Ethics Committee (ref 2011/73).

### SK152

SK152 was excavated in June, 1985, from grave 523E in a cemetery on the north side of a church in Trondheim, Norway (Figure 1A). The skeleton rested on a wooden plank (symbolic coffin) (Figure 1B). According to the daybook written in the field, the grave was filled with "grey, sandy loam, wood chips" which likely represents a mixture of minerogenic soil and waste from wood-building activities in the nearby, densely built-up area. The grave itself was covered with a wooden chip layer derived from later building activity in the area. The stratigraphical information indicates that the environment surrounding SK152 was anoxic, acidic, and waterlogged.

#### *Preservation*

The overall surface preservation of bone elements was very good, grade 1 according to the seven-category grading system defined by McKinley [76]. The skeleton was over 90% complete, with almost all bone elements present. Only minimal localized areas of bone damage due to post mortem disturbance were noted, and a small degree of fragmentation. Discoloration due to burial taphonomy was noted, but this was characteristic of the whole assemblage and did not impact on the visual inspection of bone surface.

#### *Age and sex*

Skeleton 152 was assigned to the young adult category (19-24 years) [77] on the basis of dental eruption and dental wear pattern, degeneration of the pelvis and fusion of epiphyses. Standard techniques for osteological sex determination [78] indicated that SK152 was female, and this conclusion was confirmed by the ratio of sequences aligning to the Y chromosome relative to the total number of sequences aligning to both sex chromosomes [79]. The mtDNA haplogroup was H4a.

#### *Origins*

The oxygen composition in enamel apatite carbonates from the first (M1) and third (M3) molars yielded M1 and M3 $\delta^{18}O_{Carbon}$ values of 21.13‰ and 24.72‰ on the VSMOW scale, respectively. The $\delta^{18}O_{Carbon}$ data was converted to $\delta^{18}O_{water}$ (oxygen composition in water/precipitation) by using Equation 6 from citation [80] as modified by citation [81], to yield $\delta^{18}O_{water}$ values of −15‰ to −16.5‰ for M1 and −9.3 ± 1‰ for M3 [14]. The values from the first molar suggest that SK152 was born inland in northern Scandinavia or in the north-western regions of Russia, whereas the values from the third molar indicate that she arrived in Trondheim in her childhood years [14].

## METHOD DETAILS

### Dating of SK152

The burial of SK152 was dated by two alternative methods (Figure S1). Grave 523E was part of sub-phase FN7/level II within an extension of the cemetery, corresponding to a building phase when the cemetery had recently been extended northward, and covered older/earlier buildings from sub-phase FN6. The construction of FN6 began ca. 1150 CE. Sub-phase FN7 was initially dated to "mid 1200s" with the aid of ceramics, coins, $^{14}$C-dating and archaeological lead artifacts (Reed in citation [82], p. 192), but this estimate was later revised to 1175-1225 CE on the basis of additional dendrographic dating (Figure 23 in citation [83]). Accelerator Mass Spectrometry (AMS) dating was performed at DirectAMS and Oxford Radiocarbon Accelerator Unit, yielding dates in radiocarbon years BP (Before 1950 CE) using the $^{14}$C half-life of 5,568 years. Isotopic fractionation was corrected using the $\delta^{13}$C values measured on the AMS. The quoted $\delta^{13}$C values were measured independently on a stable isotope mass spectrometer (to ± 0:3 per mil relative to VPDB). For details of the chemical pre-treatment, target preparation and AMS measurement see [84, 85]. Figure S1B shows calendar age ranges calculated by the OxCal computer program (v4.2) of C. Bronk Ramsey [48], using the 'IntCal13' dataset

[86]. These calculations support dates of 994-1052 CE with 53% likelihood and 1081-1152 CE with 42% likelihood. We summarized these results as 1,073 ± 79 years, which is the mean ± range of the lowest and highest extremes.

## Metagenomic sequencing of samples from SK152

All molecular work including pre-library amplification was conducted in dedicated aDNA clean laboratory facilities at the Centre for GeoGenetics, Natural History Museum, University of Copenhagen. All samples were collected and processed using strict aDNA guidelines. Nine sequencing libraries were prepared from the upper 3$^{rd}$ left molar root including dental pulp, the upper 2$^{nd}$ right molar dentine/cementum (200 mg) and pulp (interior root canal, 200 mg), femoral long bone (300 mg), and a mixture of mineralized dental plaque (calculus, 30 mg) taken from multiple teeth. These samples were processed in a specialized drill room within the dedicated aDNA facilities.

In the initial investigation, the entire root of the upper 3$^{rd}$ left molar was crushed, and DNA extracted according to Rohland and Hofreiter [87], with a pre-digestion step as described by Allentoft et al. [88]. The same protocol was used for the interior dental pulp. For the upper 2nd right molar, the entire tooth was removed from the mandible, and the tooth crown separated horizontally from the tooth root with a diamond-dust-coated cutting disk in a mechanical drill. The tooth root surface was then cleaned with a new cutting disk before using a small pointed drill bit to remove the interior dental pulp. DNA was extracted from this pulp as above. The remaining dentine and cementum fractions were crushed with a hammer, and also extracted. (We extracted dentine as well as cementum because this has previously maximized the yield of endogenous DNA.) The extraction protocol for this material and for dental calculus fully followed the protocols of Allentoft et al. [88] which are based on silica powder-based extraction, except that silica powder was only incubated for 1 h in the supernatant rather than the full 3 h.

DNA libraries for sequencing were prepared through blunt end ligation using NEBNext DNA sample preparation reagents (E6070) and Illumina specific adapters following established protocols. The libraries were shotgun sequenced (Table 1) in pools across 15 different lanes using Illumina HiSeq 2500 and 4000 platforms, and with a mix of 80 and 100-bp single read and 150-bp paired end chemistry. All pools submitted for sequencing contained between 5-15nM of DNA.

After initial screening, the highest proportion of *Salmonella*-specific reads were found in the upper 2nd right molar dentine/cementum (Table 1). Additional libraries were constructed from the same extracts with the same protocol, and also shotgun sequenced to increase the depth of coverage.

## Genomic assemblies from metagenomic sequences of samples from SK152

### *Genomic assemblies from metagenomic reads*

Taxonomic profiling of metagenomic reads with currently available traditional methods can yield false indications of the presence of bacterial pathogens because pathogens are over-represented in public databases relative to environmental organisms, and completed genomes of many environmental bacteria are not yet available. As a result, ecological metagenomic analyses have used *de novo* assemblies of raw reads [89] but this has not yet been applied in aDNA studies. We therefore investigated whether *de novo* assemblies could reveal the presence of bacterial pathogens within the nine metagenomic libraries from SK152, and used subsequent analysis of deamination rates to determine whether those assemblies were ancient or modern.

### *Assembly and binning*

We excluded reads within the metagenomic libraries that mapped to the human genome according to BWA [49]. The non-human reads were co-assembled into contiguous sequences (contigs) with MEGAHIT [50], using default parameters. Contigs which exceeded 20 kb were split into 10 kb fragments. All 39,016 contigs greater than 1 kb in length were clustered by CONCOCT [12] into 76 bins using both sequence composition and coverage across all samples. All of the contigs in each bin were potentially derived from a single bin-specific genomic assembly. Protein encoding sequences on these contigs were called using PRODIGAL [52], and then categorised in terms of function according to Clusters of Orthologous Groups of proteins (COGs) [90]. We previously identified 36 single copy core COGs (SCGs) that are found in a single copy in all bacterial genomes [12]. We therefore tested each of the bins for the number of copies of these 36 SCGs, and found that 11 bins largely represented MAGs (unique metagenome assembled genomes) because they contained at least 27 SCGs (75% of 36) in a single copy.

### *Taxonomic assignments*

We constructed a phylogenetic tree of the 36 SCGs from 1,755 reference genomes plus the 11 MAGs. The reference genomes consisted of one representative from each bacterial genus and each archaeal species with complete genome sequences in NCBI. Each SCG was aligned separately using MAFFT [53], and overhangs were trimmed with TRIMAL [54]. Where exceptional MAGs contained multiple SCGs, we chose one of the sequence variants at random. The concatenated SCG alignments were used to construct a single tree with FASTTREE [55]. For taxonomic assignments, we identified the ancestral node containing each MAG plus one or more reference genomes, and the most frequent taxonomic designation among the neighboring reference genomes was assigned to that MAG (Concoct 3, see Key Resources Table). C72 was initially assigned to *Mogibacterium timidum*, but only few related reference genomes existed in NCBI. We therefore compared C72 against the rMLST database [91] which contains > 200K genomes representing over 6,000 bacterial species. To this end, the translated amino acids of representative alleles in the rMLST database were mapped onto C72 using TBLASTN. This identified the positions of 46 of the 53 ribosomal genes. Those 46 ribosomal genes were then compared against all genomes in the rMLST database using TBLASTN. Analyses of both concatenated sequences and gene-by-gene analyses showed that C72 was more similar to *Eubacterium sulci* (74% identity) than *Mogibacterium timidum* (70%). We therefore re-assigned C72 to *Eubacterium.*

### Calculation of deamination rates

BWA was used to map the metagenomic reads onto all contigs within each MAG, as well as to the human genome hg38 and Paratyphi C RKS4594. We then used mapDamage 2.0 [56] to separately characterize the DNA damage associated with each of the 13 organisms. The posterior mean estimates and standard deviations for δS, the single-stranded deamination rate for all genomes and MAGs (Concoct 3, see Key Resources Table) are depicted graphically in Figure 1D.

### Interpretation of sources of MAGs

According to Kistler *et* al. [15], nine of the MAGs likely reflect recent soil contamination because their 5′-single-stranded deamination rates were low (DNA damage < 0.22) versus values of 0.47 for the human DNA and 0.9 for Ragna (Concoct 4, see Key Resources Table). C69 was unambiguously classified as *Methanosphaerula palustris*, a archaeal methanogen which is associated with acidic peat bogs [92], and belongs to a group of hydrogenotrophic methanogens with no human-associated relatives. Similarly, C40 and C66 were assigned to the environmental sulfate-reducing species *Desulfatiglans anilini* and *Desulfomonile tiedjei,* respectively, but not to human-associated sulfate reducers, such as *Desulfovibrio* species [93].

Two other assembled genomes exhibited high levels of DNA damage. C72, a novel species of *Eubacterium*, may have been a major component of a biofilm associated with periodontal disease [16]. In support of this role, C72 was recovered almost exclusively from dental calculus whereas reads from the putative environmental taxa were present in multiple sources (Concoct 5, see Key Resources Table). C18 belongs to *Acidovorax*, which is associated with plant pathogens [17], and may have been introduced with the wood chips that covered the skeleton.

## Genotyping by EnteroBase

For selected genera, EnteroBase (see Key Resources Table) automatically assembles genomes from all publicly available Illumina short reads as well as from short reads that are uploaded by users [18]. *De novo* assemblies of Illumina reads are superior to reference based mapping because they recover accessory genomic regions which are not necessarily present in the reference genome. However, all current assemblers yield a certain proportion of false base calls which do not necessarily represent the consensus of all reads. EnteroBase therefore implements a post-assembly pipeline to call SNPs based on the consensus. In brief, the ends of sequenced reads with base quality less than 5 are removed (trimmed) using Sickle [74, 94] and assembled into contigs using SPAdes 3.5 [57] with the parameter '–only-assembler' and k-mers equal to 0.3, 0.5, 0.7 and 0.9 of the average read lengths. To validate the consensus call for each base in the assemblies, the original trimmed, sequenced reads are mapped back to the corresponding assembled contigs using BWA [49], and analyzed with SAMtools/BCFtools 1.2 0.7.12-r1039 [95]. The quality scores for consensus calling for each base are stored together with the assemblies in standard FASTQ format. Finally, assembled contigs are assigned taxonomic labels using Kraken [42] in order to exclude potential contamination from other genera.

Assemblies that pass internal quality criteria, including a mean coverage of ≥ 20-fold are automatically genotyped by multiple multi-locus sequence typing (MLST) schemes into Sequence Types (STs) consisting of unique allele numbers for each genetic locus. Details of these schemes are available on the Help pages at EnteroBase. For *Salmonella*, these MLST schemes currently include a 7 housekeeping gene legacy scheme [3] and rMLST based on 51 ribosomal proteins as defined by Jolley et al. [91], and whose reference allelic sequences are maintained at the primary rMLST database at Oxford University [91]. EnteroBase also calculates alleles for wgMLST based on a pan-genome of 21,065 genes. In brief, the wgMLST scheme encompasses unique sets of homologs with ≥ 70% pairwise amino acid similarity over 50% of their length, and which were defined on the basis of 537 diverse, high quality *Salmonella* genomes. These homolog sets are usually either absent from individual *Salmonella* genomes, or are present only in a single copy. cgMLST (core genome MLST) V2 consists of that subset of 3,002 loci from the wgMLST scheme which met the following conditions for the 3,144 rMLST STs in *S. enterica* that had been defined by May, 2016: presence in ≥ 98% of genomes; an intact reading frame in ≥ 94%; and of unexceptional genetic diversity. For several homolog sets, more than one copy is occasionally present per genome. When rare genomes contain two or more copies of any wgMLST or cgMLST locus, that locus is scored internally as duplicated for that genome, and is hidden to public access for that genome.

### Species trees of S. enterica subspecies I

Most aDNA analyses have been performed with genetically monomorphic pathogens with only limited genetic diversity, and where recombination is rare. However, typical species containing bacterial pathogens, such as *S. enterica,* are usually much more genetically diverse, and commonly also undergo recombination. Until now a reliable and publicly available phylogenetic topology of *S. enterica* genomes did not exist, and it was not readily possible to accurately assign individual metagenomic reads to individual lineages [96]. We therefore assembled a collection of genomes that represent the entire genetic diversity of 50,000 genomes of *S. enterica* subsp. I by choosing one random representative of each of the 2,964 rMLST STs from that subspecies that were present in EnteroBase (May, 2016) (*Salmonella* supertree 1, see Key Resources Table). These genomes can be accessed at the Enterobase Workspace rST Representatives (see Key Resources Table). We performed super-tree analyses on the sequences of the 3,002 core genes in cgMLST v2 from these 2,964 representatives. The sequences of each core gene were aligned using Mafft [53]. These alignments were used to calculate species trees by two different algorithms. i) A Maximum Likelihood tree of the concatenated, aligned sequences of all core gene sequences (2.8 Mb) (*Salmonella* supertree 3, see Key Resources Table) was generated using a GTRCAT model in RAxML v8.2.4 [43] (Figure 2A). ii) Maximum Likelihood trees were also generated by RAxML for each of the 3,002 core gene alignments, and these 3,002 gene trees were processed by Astrid [44]. Astrid is a recently developed coalescent-based method for species tree estimation that is statistically consistent under the multi-species coalescent, can account for incomplete lineage sorting, is one of the most accurate currently available methods, and can handle large amounts of data. Because Astrid only infers topology,

we derived the branch lengths from sequence distances within the concatenated alignment using RAxML with a GTRGAMMA model (*Salmonella* supertree 2, see Key Resources Table). The topologies of the two trees were identical near the root and toward the tips, and at 75% of the intermediate branches. These trees were used for phylogenetic placement (MGPLACER) [58] of the metagenomic reads from SK152. We also compared the results from ASTRID with those from a separate species tree (data not shown) calculated with the help of a GPU compilation of ASTRAL-II [97] by S. Mirarab and T. Warnow. The species trees generated by ASTRAL and ASTRID were almost indistinguishable (data not shown). In all three species trees, two genomes from the rare serovar Birkenhead clustered unambiguously with the Para C lineage, and were used as an outgroup for rooting trees in further analyses (Figure 2).

## Sources of genomes from the Para C Lineage

The species trees identified 100 genomes in Enterobase that were in the Para C lineage, or closely related to it. Most of these genomes were very recent, epidemiological isolates from the UK (PHE) and USA (FDA), but a number of them were from strains sequenced by the Wellcome Trust Sanger Centre that had been tested by legacy MLST [3] or were from the historical Murray collection [98]. In order to ensure that we spanned diverse sources and dates, we cultivated 119 additional old isolates from the historical, global collection at the Institut Pasteur, Paris, and also sequenced their genomes.

Total DNA was extracted with a Maxwell 16 cell DNA purification kit (Promega, Madison WI), in accordance with the manufacturer's recommendations. Libraries were constructed using the Nextera XT kit (Illumina) with the following modifications. The initial tagmentation reaction was performed using 2 μl of 0.7 ng/μl of template DNA and 2/5 of the specified volume for other reagents, resulting in a volume of 10 μl after neutralization. For the PCR step, 25 μl of 2x Extensor Hi-Fidelity PCR master mix (Thermo Scientific), 5 μl of each index primer (4 μM) and 5 μl of sterile distilled water was added to the tagmentation reaction. The standard PCR reaction was extended by an extra 3 cycles, and the extension step was lowered from 72°C to 68°C. The libraries were purified using 25 μl of Ampure XP beads (Beckman Coulter) with two 200 μl washes with 80% ethanol before elution in 30 μl of RSB from the Nextera XT kit. Libraries were quantified using the Qubit dsDNA HS Assay Kit (Thermo Scientific), and diluted to 3.2 ng/μl (approx. 8 nM). Pooled libraries (40 samples per run) were denatured following the Illumina protocol, and 600 μl (approx. 20 pM) was loaded onto a MiSeq V2 −500 cycle cartridge (Illumina), and sequenced on a MiSeq to produce FASTQ files of short reads.

The short reads were uploaded to EnteroBase and to the Short Read Archives. Their metadata and genome properties are summarized together with other genomes of the Para C lineage in Table S5 and can be examined interactively in the EnteroBase public workspace Para C lineage (see Key Resources Table).

## *Salmonella* short reads from metagenomic aDNA

### *Initial analyses*

Metagenomic reads from 33 human skeletons from Trondheim, Norway were screened using KRAKEN [42] with its default genomic database (Workflow, see Key Resources Table). KRAKEN scored 304 metagenomic reads from skeleton SK152 as being *Salmonella,* and did not score any reads from the 32 other skeletons as *Salmonella* or other potentially invasive bacterial pathogens. However, the taxon assignments provided by KRAKEN can be problematical when reads are present at low concentrations [99]. In our hands, the KRAKEN analyses of SK152 also failed to identify seven of the eleven taxa that yielded assembled MAGs with CONCOCT (Concoct 6; Concoct 7, see Key Resources Table). We therefore tested the reads from SK152 that had been identified as *Salmonella* by *ad hoc* BLASTN alignments to a custom database based on 91,000 bacterial, archaeal, and viral genomes from the non-redundant nucleotide database in GenBank, and confirmed that 266 (87%) scored as *S. enterica*. MGPlacer [58] was then used to iteratively map those *Salmonella*-specific metagenomic reads onto a core SNP phylogeny based on one genome from each of the 20 *Salmonella* serovars represented in RefSeq (May, 2016). All *Salmonella*-specific reads mapped on the branch leading to serovar Paratyphi C strain RKS4594.

### *Ragna-specific reads in metagenomic libraries*

Reads were pre-processed using BBMERGE and BBDUK2 from the BBMAP package [60]. Sequences specific to the Ragna genome were then identified by alignment against two sets of reference genomes using BOWTIE2 [61]: i) The "ParaC" set, consisting of all 219 modern genomes in the Para C lineage and ii) the "outgroup" set, consisting of representatives of all (4,441) non-*Salmonella* bacterial genomes in RefSeq, plus the human genome hg38. We assigned sequencing reads to Ragna if they yielded equal or better alignment scores to the ParaC set than to the outgroup, and differed in sequence from the most similar genome in the ParaC lineage by no more than 4%. Potentially duplicated reads were removed by retaining only one Ragna-specific read sequence when multiple identical reads were identified. The quality of the Ragna-specific reads are summarized in Figure S2.

### *SNPs in the Ragna genome*

The Ragna-specific, unique reads were aligned against Paratyphi C reference genome RKS4594 using BOWTIE2 [61] with the end-to-end option and analyzed using SAMTOOLS/BCFTOOLS 1.2 [62]. In order to exclude potential spurious SNPs due to deamination (Figure S2C), the consensus base was only called on sites that were covered by at least two reads with a consensus base quality $\geq$ 10 and which were located at least 5 bases from either read end (Figure S2B). Exceptionally, SNP calls with single coverage matching these criteria were included if the same call was provided by previously excluded 5′ and 3′ ends of reads.

### *Reference based SNP calling in modern genomes*

Assemblies within the Para C lineage plus Birkenhead (Figure S4) were aligned against RKS4594 using LAST [63], and SNPs from these alignments were filtered to remove regions with low base qualities (Q < 10) or ambiguous alignment (ambiguity $\geq$ 0.1). Sites were also removed if they aligned with $\geq$ 95% identity to disperse repetitive regions that were longer than 100 bp (BLASTN),

overlapped with tandem repeats (TRF) [64], or overlapped with CRISPR regions (PILER-CR) [65]. Of the remaining SNPs, 61,451 were called in $\geq$ 95% of the genomes from the ParaC Lineage plus Birkenhead, and were retained as core genomic SNPs (ParaC genome SNPs 2, see Key Resources Table).

#### *Reconstruction of the mutational phylogeny of the Para C lineage*

An initial phylogeny (ParaC genome SNPs 3, see Key Resources Table) was calculated on all 61,451 core SNPs from the Para C lineage plus Birkenhead using RAxML v8.2.4 [43] under a GTRCAT model with Stamatakis ascertainment correction for invariant sites [100]. SNPs were assigned using RecHMM [25] onto branches in that phylogeny using a maximum likelihood method with a symmetric transition model [101]. The results indicated that 21,071 genomic sites had suffered substitution events along the branches leading to serovars Paratyphi C, Typhisuis and Choleraesuis, of which 406 sites (1.9%) were mutated on multiple independent occasions (homoplasies) resulting in a total of 21,849 substitution events (ParaC genome SNPs 1; ParaC genome SNPs 2, see Key Resources Table). RecHMM also identified 1,602 SNPs that were clustered, which are the hallmarks of homologous recombination. After excluding the recombinational SNPs, the remaining 20,247 core mutational SNPs were used for further phylogenetic analyses. The RAxML phylogeny based on these mutational SNPs (Figure 2B) differed only in branch lengths but not in topology from the initial RAxML phylogeny based on all core SNPs. The mutational phylogeny was also used for additional analyses of gene gain and loss and numbers of pseudogenes (Figure 3; Para C Pan-genome 1, see Key Resources Table) and ancestral reconstruction (Ancestral reconstruction 1, see Key Resources Table).

In separate experiments, recombinant SNPs on the branches to serovars Paratyphi C, Typhisuis and Choleraesuis were called by the alternative programs ClonalFrameML [102] and Gubbins [103] and compared with the calls by RecHMM. 20,023 SNPs were scored as mutational and 1,407 SNPs as recombinational by all three programs, but the programs differed in their assignments for 398 SNPs (1.8%) (ParaC genome SNPs 4, see Key Resources Table). In contrast, the three programs differed dramatically in their assignments of SNPs on the branches to Lomita and Birkenhead, and we did not attempt to subdivide those SNPs into recombinational versus mutational.

### Date estimates for Paratyphi C and related serovars

Date estimates based on Bayesian phylogenetic analysis are summarized in Table S4.

#### *Bayesian phylogenetic approach*

Core mutational SNPs were analyzed using BEAST v1.8.3 [45] with the GTR+G model of nucleotide substitution, with a discrete gamma distribution and six rate categories to account for rate heterogeneity across sites. To account for SNP ascertainment bias, we applied a correction that incorporated the nucleotide frequencies across all of the constant sites. For computational tractability, we analyzed two subsamples each from the Para C lineage without Lomita; serovar Paratyphi C serovar (including Ragna); and modern crown serovar Paratyphi C strains (excluding Ragna). Each subsample comprised 50 random genomes, except that the Ragna sequence and the sole genome in the CS-3 cluster were deliberately included when present in the test set. Sub-samples including the Ragna genomes were run with either of two dates for Ragna (1200 CE from archaeological dates and 1073 CE from AMS calibrated ages). Thus, our Bayesian phylogenetic analyses involved 12 datasets (Date estimation 3, see Key Resources Table). The raw trees are presented in Date estimation 4 (see Key Resources Table).

Posterior distributions of parameters, including the MCC tree, were estimated using Markov chain Monte Carlo sampling. Samples were drawn every 5,000 steps over a total of $5 \times 10^7$ steps, with the first 10% of samples discarded as burn-in. Marginal likelihoods were used to compare two clock models (strict clock and uncorrelated lognormal relaxed clock) and three demographic models for the tree prior (constant population size, Bayesian Skyride coalescent [104], and birth-death process with serial sampling [105]). Marginal likelihoods were estimated using stepping-stone sampling [106], with 20 path steps and a chain length of $10^6$ per path step, and the most likely model combinations are indicated in bold print in Date estimation 3 (see Key Resources Table). This table also presents the median MRCA estimates plus 95% credible intervals based on all six model combinations for the Para C lineage, Paratyphi C including Ragna and the Paratyphi C modern crown lineage (without Ragna). The optimal model for Paratyphi C including Ragna was the strict clock model with constant population size. The optimal model for the ParaC Lineage was UCLD (uncorrelated relaxed clock) model with constant population size. To ensure that these results did not represent outliers due to limited numbers of runs, we also tested eight additional subsamples using these optimal models and assuming a date for Ragna of 1200 CE. The median values for their estimates of tMRCAs and mutational clock rates and the corresponding 95% confidence intervals are summarized in Table S4.

#### *Temporal structure*

It is crucial to verify the existence of temporal structure in datasets used for date estimation. We therefore tested the reliability of the BEAST analyses by date randomization tests [38] which analyze multiple, date-randomized replicate datasets after randomly reassigning the ages of the sequences. Datasets are considered to have strong temporal structure when the 95% credible interval of the rate estimate from the original data does not overlap with those of the rate estimates from the date-randomized replicates [107]. We used 10 date randomizations for each of the initial two sub-samples, with and without Ragna, and found strong temporal signals within all sub-samples of modern genomes analyzed with BEAST (Date estimation 1, see Key Resources Table).

#### *Root-to-tip regression*

We also used an alternative method to estimate temporal signal for dating on the non-recombinational SNPs in 116 Paratyphi C genomes and the 205 Para C lineage genomes with known collection dates, including Ragna but excluding Lomita. This consisted of a regression of root-to-tip distances with TEMPEST [37] (previously designated PATH-O-GEN) in which the degree of within-lineage

rate heterogeneity is evaluated by calculating the correlation coefficient. A regression based on one aDNA genome (Ragna) plus multiple modern genomes is basically a regression line based on two points. We therefore also included two additional aDNA genomes, Tepos-14 and Tepos-35 from 1,545 CE in Mexico [19]. The results (Data estimation 2, see Key Resources Table) showed a clear temporal signal in both datasets, with an extrapolation to the root which was similar to the tMRCAs found with BEAST. The Paratyphi C dataset yielded $R^2$ of 0.38, which confirms the existence of temporal signal. The Para C lineage only had an $R^2$ of 0,04 indicating that TEMPEST is not suited for the analyses of these data.

### Inferring ancestral geographic sources

A subtree of genomes with geographic source information was extracted from the Para C lineage phylogeny plus the outgroup Birkenhead with the ETE3 Python package [66]. (Lomita is lacking from this subtree because its geographic source is uncertain). The ancestral geographic states of internal nodes in the subtree were inferred by three independent algorithms: i) the Maximum Likelihood comparison [108] in BayesTraits [47]; ii) the Markov chain Monte Carlo (MCMC) approach [109] in BayesTraits; and iii) Bayesian Binary MCMC in RASP [75]. All three algorithms indicated a European origin for individual nodes within the Para C lineage (Ancestral reconstruction 1; Ancestral reconstruction 2, see Key Resources Table), and this conclusion was not affected by the inclusion or absence of Ragna. A European origin of the entire Para C lineage was also strongly supported by BayesTraits MCMC and RASP, and moderately supported by BayesTraits ML (Ancestral reconstruction 2, see Key Resources Table).

### Reconstruction of the pan genome of the Para C lineage

A wide collection of annotated reference genes was collected from three sources: 1) all 21,065 unique sets of homologs in the wgMLST scheme in Enterobase; 2) published annotations for the complete genomes of Choleraesuis strains C500 [110], SC-B67 [111] and Paratyphi C strain RKS4594 [112]; 3) PROKKA [67] annotations of all draft genomes in EnteroBase. All these reference genes were grouped into 29,436 gene clusters using UCLUST [68]. In order to obtain sets of homologous regions that cover $\geq$ 50% of the length of the centroid sequences with $\geq$ 70% nucleotide identity, the centroid sequence from each cluster was aligned with all the modern genomes in the Para C lineage using BLASTN. Overlapping paralogs between the homolog sets were identified through the same iterative methodology used to construct the entire *Salmonella* wgMLST scheme. After removal of paralogs, the remaining 6,665 homolog sets were treated as the pan genes of the Para C lineage (ParaC Pan-genome 4; ParaC Pan-genome 5, see Key Resources Table).

#### *Reconstruction of pan genome synteny*

Synteny was reconstructed as described in Supplemental Materials of Zhou et al. [26], namely through constructing and traversing a graph of assembled sequences from genes in the Para C lineage pan-genome. First, the graph was seeded with one node for each unique, single copy gene. The connections between nodes were weighted using the following criteria: 1) Edges connecting pairs of genes that were co-located on a single contig received maximal weighting. 2) Edges received intermediate weighting that connected two genes at the ends of distinct contigs that were however linked by read-pairs that straddled both contigs. This intermediate weighting was 2*(number of read-pairs joining the two contigs)/(total number of unpaired reads at the ends of contigs). 3) Pairs of genes which did not co-locate according to either of these two criteria were not assigned to a common edge.

CONCORDE [51] was used to find the shortest possible path that visited all the nodes in the graph, which equates to the most likely gene order within the Para C lineage. Ambiguous paths were inspected manually to identify duplicated genes and collapsed repeats, because these are usually associated with prophages or plasmids. Such connections were then broken manually and re-joined as appropriate. Finally, in order to reconstruct the synteny of the pan genome of the Para C lineage, all repetitive genes were inserted into the gene order according to their location within the assemblies (Figure 2B and ParaC Pan-genome 4, see Key Resources Table). A total of 227 genomic islands (ParaC Pan-genome 1; ParaC Pan-genome 7, see Key Resources Table) were identified as continuous blocks of gene gain/loss in the pan genome of the Para C lineage.

The conservation of genes within the pan genome across the Para C lineage was illustrated using Anvi'o [73] (Figure 2B), and can be examined in detail in a publicly accessible, interactive EnteroBase version of Anvi'o (Pan-genome, see Key Resources Table). The synteny of pan genes was enforced within the Anvi'o rendering by help of an artificial guide tree based on the gene order in ParaC Pan-genome 4 (see Key Resources Table), where each sequential gene bifurcates from the previous gene at a constant distance. (That artificial tree was deleted from the printed version in Figure 2B). The figures are dependent on manually generated input files based on the annotations and sub-divisions that are indicated in ParaC Pan-genome 4 (see Key Resources Table). The input files also include the frequencies of pan genes in all sub-lineages, the locations of major genomic islands, as well as additional metadata. These input files can also be downloaded from the interactive EnteroBase version.

#### *Gain and loss of pseudogenes and intact genes*

The gene alignments used to reconstruct the pan genome of the Para C lineage were screened for potential gene disruptions. Genes with stop codons or frameshifts anywhere in the coding regions were scored as pseudogenes. Pseudogenes and genes that had been gained or lost were assigned to ancestral states in the core SNP phylogeny (ParaC Pan-genome 1, see Key Resources Table) by the Maximum Likelihood algorithm [101] as implemented in the R function ACE (Ancestral Character Estimation) within the APE (Analyses of Phylogenetics and Evolution) package (ParaC Pan-genome 6; ParaC Pan-genome 7, see Key Resources Table) [69]. These ancestral states were used to infer the phylogenetic branches on the SNP tree of the Para C lineage on which 3,125 genes or parts of genes and 1,251 pseudogenes were independently gained or lost (Figure S3; ParaC Pan-genome 1; ParaC Pan-genome 3, see Key Resources Table). The relative frequencies of gain/loss of intact genes and pseudogenes versus time were estimated as

described [25]: time (years) was calculated from the core SNP phylogeny by dividing branch lengths by the median SNP substitution rate estimated by BEAST ($7.9 \times 10^{-8}$ substitutions per site per year; Table S4). The 95% confidence intervals for these frequencies of pseudogenes and gene gain/loss were inferred from 1,000 separate bootstrap re-samplings of pseudogenes or genomic islands. The data and script used for these purposes can be found in TemporalFreq.py and Data (see Key Resources Table).

#### *Identification of bacteriophages*

Prophages in modern genomes of the Para C lineage were identified using the online API in PHASTER [70]. The identified prophages were taxonomically labeled according to the similarities of their major capsid protein (MCP) to the clusters of bacteriophages described by Casjens et al. [113]. Further manual taxonomic refinements were based on comparisons of all genes in each prophage against those in the published phage clusters. Prophage names are the same as the previously described phages which they most closely resembled, except for GI28, which has no known close relatives, and SPC-P1, which had been previously designated under that name in the annotation of Paratyphi C RKS4594 [28].

#### *Annotation of genomic islands*

BLASTN was used to identify known genomic islands within the pan genome as tight clusters of genes in which $\geq$ 60% of the sequences aligned with a previously described SPI/SGI sequence with $\geq$ 80% nucleotide identity. The genomic sequences used as references for SPI-1 through SPI-12 were downloaded from PAIDB [40], as were SGI-1 and SGI-2. SPI-13 through SPI-21 were obtained from citations [114–118].

Genes in the pan genome were also screened for an association with mobile elements that are listed in ISFinder (IS elements) [41], PlasmidFinder 1.3 (incompatibility groups of plasmids) [71] and CONJScan-T4SSScan (relaxases and key components of type IV secretion systems) [72]. Strong matches were annotated according to those resources (Figure 2B; ParaC Pan-genome 4; ParaC Pan-genome 7, see Key Resources Table). These analyses identified 227 genetic islands belonging to 127 distinct categories in which one or more genes were gained or lost (ParaC Pan-genome 3, see Key Resources Table).

## QUANTIFICATION AND STATISTICAL ANALYSIS

No statistical methods were used to predetermine sample size. The experiments were not randomized except for Bayesian analyses and the investigators were not blinded to allocation during experiments and outcome assessment.

## DATA AND SOFTWARE AVAILABILITY

Genomic reads for 119 strains from the Institut Pasteur collection have been deposited in the NCBI short read archive under project accession GenBank: PRJEB19916. All genomes referred to here are available from the publicly available workspaces entitled "rST representatives" (2,964 genomes) and "Para C lineage" in the *Salmonella* database of EnteroBase (http://EnteroBase.warwick.ac.uk). Interactive versions of Figure 2 and Figure S4 can be found at https://enterobase.warwick.ac.uk/anvio/public/zhemin/ParaC_pangenome and https://enterobase.warwick.ac.uk/anvio/public/zhemin/ParaC_SPI6. Additional data, figures, and tables are permanently stored at the University of Warwick and non-human metagenomic data at https://sid.erda.dk/wsgi-bin/ls.py?share_id=E56xgi8CEl and can also be accessed using the links provided in the Key Resources Table.